STICH – A Hierarchical Clustering Algorithm

Maribel Yasmina Santos, Adriano Moreira and Sofia Carneiro

Department of Information Systems, University of Minho, Portugal
{maribel, adriano, sofia}@dsi.uminho.pt

Abstract. Clustering has been widely used to find homogeneous groups of data in datasets while looking at some specific metric. Several clustering techniques have been developed, each one presenting advantages and drawbacks to specific applications. This work addresses the development of a clustering technique for the creation of Space Models – STICH (*Space Models Identification Through Hierarchical Clustering*). Space Models are divisions of the space in which the elementary regions are grouped according to their similarities with respect to a specific indicator (value of an attribute). The identified models, which are formed by sets of clusters, point out particularities of the analysed data, namely the exhibition of clusters with outliers, regions which behaviour is strongly different from the other regions analysed. The results achieved with STICH and with the well known *k-means* algorithm are compared, allowing the validation of the work developed so far in STICH.

1 Introduction

Maps are extensively used by humans as a mean to support data visualization. The geometry explicit in each map was defined for or with a specific purpose, being sometimes later used in applications for which it was not conceived. One example are maps representing administrative subdivisions of the geographic space. Administrative subdivisions inside a country – parishes, municipalities, and districts, for example – are defined following specific criteria and not following a natural division of the space. They are later used as a tool for data analysis, even when they are not the most suitable resource for that.

This work addresses the problem of creating a natural division of the space. We start by defining Space Models as a new space geometry that is created to emphasize the particularities of the data used in its creation. This means that a Space Model is defined by sets of elementary regions, grouped accordingly to the similarities that exist between them. Elementary regions inside the same set are similar regarding a specific characteristic, and regions in different sets are as dissimilar as possible.

Space Models are of great importance for the analysis of indicators (environmental or social, for instance) associated to regions and to understand the main differences between these regions. This is the objective of the EPSILON (*Environmental Policy via Sustainability Indicators on a European-wide NUTS III Level*) project[1], in which

[1] A project founded by the European Commission through the IST program (contract IST-2001-32389).

sustainability indicators are analysed in order to identify regions with similar behaviour, and regions that exhibit different trends in data. The final objective of the EPSILON project is to contribute for a sustainable development across Europe. This project contributes to the better understanding of the European Environmental Quality and Quality of Life, by delivering a tool aimed to generate environmental sustainability indices at NUTS-III level.

The approach undertaken for the creation of Space Models is based on clustering techniques, and allows the creation of Space Models that point out different levels of aggregations and particularities in the analysed data.

STICH (*Space Models Identification Through Hierarchical Clustering*) assumes several principles in the identification of Space Models, namely:

- Space Models must be created from the values available for the indicators, and no constraints must be imposed in the creation process.
- The created Space Models must be the same, independently of the order by which the available data is processed.
- Space Models can include regions (clusters) of different shapes and sizes.
- Space Models must be independent of specific domain knowledge, like the specification of an initial number of regions.

This paper is organised as follows: Section 2 introduces clustering, a data mining technique, and presents two well-known types of clustering algorithms, *k-means* and *k-nearest neighbour*. STICH, the clustering technique proposed in this work for the creation of Space Models is also described in section 2. Section 3 presents the results achieved with *k-means* and with STICH, the proposed clustering algorithm. Section 4 evaluates the obtained results, pointing out the advantages of STICH. Section 5 concludes with some remarks about the presented work.

2 Clustering: a Data Mining technique

Clustering is a discovering process that groups a set of data objects in a way that maximises the similarity between the objects inside a cluster, and minimise the similarity between different clusters [1] [2]. It is considered a data mining technique and an unsupervised learning technique since the user has no influence in the discovery process [3].

Two of the well-known types of clustering algorithms are based on partitioning and hierarchical methods [1]. Partitioning algorithms identify a partition of a database, assigning its n objects to a set of k clusters, where k is an input parameter. Each cluster is represented by the gravity centre (centroid) of the cluster (*k-means*) or by one of the objects of the cluster located near to its centre (*k-medoid*) [4]. In these algorithms, each object is assigned to the closest representative cluster (the cluster for which the distance between the object and the centroid is minimum).

Hierarchical clustering algorithms create a hierarchical decomposition or composition of a given set of data objects. They can be agglomerative or divisive, based on how the hierarchical clustering process is performed. In hierarchical

methods, once a step (merge or split) is done, it can never be undone, which does not allows the correction of erroneous decisions [1].

The following sections introduce the *k-means* and *k-nearest-neighbour* algorithms, as examples of partitioning clustering methods and hierarchical clustering methods respectively, and STICH, the proposed algorithm for the creation of Space Models through hierarchical clustering techniques.

2.1 The *K-means* algorithm

Partitioning-based clustering algorithms such as *k-means* attempt to break data into a set of *k* clusters. This partition tries to optimize a given criterion and assumes that clusters are hyper-ellipsoidal and of similar sizes [5] [6].

The *k-means* algorithm takes as input a parameter *k* that represents the number of clusters in which the *n* objects of a data set will be partitioned. The division obtained tries to maximise the *Intracluster* similarity (a measurement of the similarity between the objects inside a cluster) and minimise the *Intercluster* similarity (a measurement of the similarity between different clusters). This similarity is measured looking at the mean value of the objects in a cluster, which represent the centre of gravity (centroid) of the clusters.

Given the input parameter *k*, the *k-means* algorithm works as follows [1]:

1. Randomly selects *k* objects, each of which initially represents the cluster centre or the cluster mean.
2. Assign the remaining objects to the cluster to which each record is the most similar, based on the distance between the object and the cluster mean.
3. Compute the new mean (centroid) of each cluster.

After the first iteration, each cluster is represented by the mean calculated in step 3. This process is repeated until the criterion function converges. The *squared-error* criterion is often used, which is defined as:

$$E = \sum_{i=1}^{k} \sum_{j=1}^{l} o_j \in C_i \left| o_j - m_i \right|^2 \quad (1)$$

where E is the sum of the square-error for the objects in the dada set, l is the number of objects in a given cluster, o_j represents an object, and m_i is the mean value of the cluster C_i. This criterion intends to make the resulting k clusters as compact and as separate as possible [1].

The *k-means* algorithm has as disadvantages the necessity for users to specify *k* in advance and the fact that it is not suitable for discovering clusters with nonconvex shapes or clusters with very different sizes [1].

In order to facilitate the comprehension of the *k-means* algorithm and its main differences to the *k-nearest-neighbour* and the STICH algorithms presented later, this description proceeds with an example of use of *k*-means in the identification of clusters in a dataset.

Suppose that a set of 9 objects needs to be clustered in 3 regions. Let *D = {3, 5, 8, 12, 14, 18, 20, 22, 24}*. According to the algorithm described above (depicted in Fig. 1), in the first iteration three objects are randomly selected as the three seeds

(centroids) of the clusters ($C_1 = 3$, $C_2 = 5$, $C_3 = 8$). The 9 objects are assigned to the available clusters accordingly to their distances to the cluster centre. After this distribution (end of 1st iteration), the mean of each cluster is calculated, considering all the objects inside it. The overall square-error is calculated and in this 2nd iteration the difference between each object and its cluster mean is identified, in order to verify the redistribution of the objects by the clusters. This process can be followed in Fig. 1, where it is possible to verify that six iterations were needed for the identification of the final distribution of the objects by the clusters (in the last iteration no redistributions were verified, meaning that the process reached the end). By the analysis of the mentioned figure it is also possible to see that the value of *E*, the squared-error criterion, varied from *198.9* at the beginning of the process to *40.7* at the end of the process.

	1						2							3						
n		*m*	*Dif. C 1*	*Dif. C 2*	*Dif. C 3*			*m*	*E*	*Dif. C 1*	*Dif. C 2*	*Dif. C 3*			*m*	*E*	*Dif. C 1*	*Dif. C 2*	*Dif. C 3*	
3	C_1	3.0	0.0	2.0	5.0	C_1	C_1	3.0	0.0	0.0	2.0	13.9	C_1	C_1	3.0	0.0	0.0	3.5	15.3	C_1
5	C_2	5.0	2.0	0.0	3.0	C_2	C_2	5.0	0.0	2.0	0.0	11.9	C_2	C_2	6.5	2.3	2.0	1.5	13.3	C_2
8	C_3	8.0	5.0	3.0	0.0	C_3	C_3	16.9	78.4	5.0	3.0	8.9	C_2	C_3	18.3	2.3	5.0	1.5	10.3	C_2
12			9.0	7.0	4.0	C_3			23.6	9.0	7.0	4.9	C_3			40.1	9.0	5.5	6.3	C_2
14			11.0	9.0	6.0	C_3			8.2	11.0	9.0	2.9	C_3			18.8	11.0	7.5	4.3	C_3
18			15.0	13.0	10.0	C_3			1.3	15.0	13.0	1.1	C_3			0.1	15.0	11.5	0.3	C_3
20			17.0	15.0	12.0	C_3			9.9	17.0	15.0	3.1	C_3			2.8	17.0	13.5	1.7	C_3
22			19.0	17.0	14.0	C_3			26.4	19.0	17.0	5.1	C_3			13.4	19.0	15.5	3.7	C_3
24			21.0	19.0	16.0	C_3			51.0	21.0	19.0	7.1	C_3			32.1	21.0	17.5	5.7	C_3
									198.9							111.8				

	4							5							6						
n		*m*	*E*	*Dif. C 1*	*Dif. C 2*	*Dif. C 3*			*m*	*E*	*Dif. C 1*	*Dif. C 2*	*Dif. C 3*			*m*	*E*	*Dif. C 1*	*Dif. C 2*	*Dif. C 3*	
3	C_1	3.0	0.0	0.0	5.3	16.6	C_1	C_1	4.0	1.0	1.0	7.0	16.6	C_1	C_1	4.0	1.0	1.0	8.3	18.0	C_1
5	C_2	8.3	11.1	2.0	3.3	14.6	C_1	C_2	10.0	1.0	1.0	5.0	14.6	C_1	C_2	11	1.0	1.0	6.3	16.0	C_1
8	C_3	19.6	0.1	5.0	0.3	11.6	C_2	C_3	19.6	4.0	4.0	2.0	11.6	C_2	C_3	21.0	11.1	4.0	3.3	13.0	C_2
12			13.4	9.0	3.7	7.6	C_2			4.0	8.0	2.0	7.6	C_2			0.4	8.0	0.7	9.0	C_2
14			31.4	11.0	5.7	5.6	C_3			31.4	10.0	4.0	5.6	C_2			7.1	10.0	2.7	7.0	C_2
18			2.6	15.0	9.7	1.6	C_3			2.6	14.0	8.0	1.6	C_3			9.0	14.0	6.7	3.0	C_3
20			0.2	17.0	11.7	0.4	C_3			0.2	16.0	10.0	0.4	C_3			1.0	16.0	8.7	1.0	C_3
22			5.8	19.0	13.7	2.4	C_3			5.8	18.0	12.0	2.4	C_3			1.0	18.0	10.7	1.0	C_3
24			19.4	21.0	15.7	4.4	C_3			19.4	20.0	14.0	4.4	C_3			9.0	20.0	12.7	3.0	C_3
			83.9							69.2							40.7				

Fig. 1. *k-means* algorithm: an example

2.2 The *k-nearest-neighbour* algorithm

Hierarchical clustering algorithms generate a set of clusters with a single cluster at the top (integrating all data objects) and single-point clusters at the bottom, in which each cluster is formed by one single data object [5].

Agglomerative hierarchical algorithms, like *k-nearest neighbour*, start with each data object at a separate cluster. Each step of the clustering algorithm merges the *k* records that are most similar into a single cluster.

The *k-nearest neighbour* algorithm uses a graph to represent the links between the several data objects. This graph allows the identification of the several *k* objects most similar to a given data item. Each node of the *k-nearest neighbour* graph represents a data item. An edge exists between two nodes *p* and *q* if *q* is among the *k* most similar objects of *p*, or if *p* is among the most similar objects of *q*. Fig. 2 represents the iterative process associated with the *k-nearest neighbour* algorithm, grouping a set of

data points in clusters with *1* and *2-nearest neighbours*. This process is presented using the example dataset provided in section 2.1, *D = {3, 5, 8, 12, 14, 18, 20, 22, 24}*.

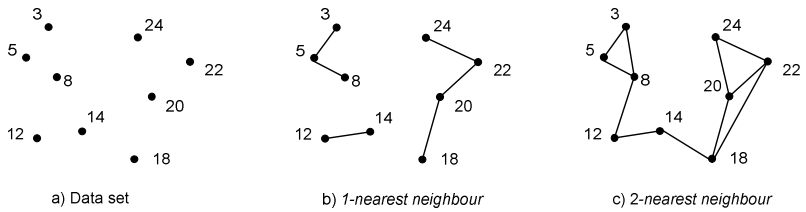

Fig. 2. *k-nearest neighbour graph*: an example

2.3 STICH

The STICH algorithm is based on the *k-nearest-neighbour* algorithm. However, the principles defined to STICH try to overcome some of the limitations of the *k-nearest-neighbour* approach, namely the necessity to define a value for the input parameter *k*. This value imposes restrictions to the maximum number of members that a given cluster can have.

STICH uses an iterative process, in which no input parameters are needed. Another characteristic of STICH is that it produces several usable Space Models, one at the end of each iteration of the clustering process.

As a hierarchical clustering algorithm with an agglomerative approach, STICH starts to assign each object in the dataset to a different cluster. Its clustering process begins with as many clusters as objects in the dataset, and ends with all the objects grouped into the same cluster.

The approach undertaken in STICH is as follows:

1. For the dataset, calculate the Similarity Matrix of the objects, which is the matrix of the distances between every pair of objects in the dataset.
2. For each object, identify its minimum distance to the dataset (the value that represents the distance between the object and its *1-nearest-neighbour*).
3. Calculate the median[2] of all the minimum distances identified in the previous step.
4. For each object, identify its *k-nearest neighbours*, selecting the objects that have a distance value less or equal to the median calculated in step 3. The number of objects selected, *k*, may vary from one object to another.
5. For each object, calculate the average *c* of its *k-nearest neighbours*.
6. For each object, verify in which clusters it appears as one of the *k-nearest neighbours* and then assign this object to the cluster in which it appear with the minimum *k-nearest neighbour* average *c*. In case of tie, that happens when one object appears in two clusters with equal average *c*, the object is assigned to the first cluster in which it appears in the Similarity Matrix.

[2] The median is used instead of the average, since the average can be negatively influenced by outliers.

7. For each new cluster, calculate its centroid.

This process is iteratively repeated until all the objects are grouped into the same cluster. The output of the several iterations of STICH represents a new Space Model that can be used for different purposes.

For each model, a quality metric is calculated after each iteration of the STICH algorithm. This metric, the *ModelQuality*, is defined as:

$$ModelQuality = \left| Intracluster - Intercluster \right| \tag{2}$$

and it is based on the difference between the *Intracluster* and *Intercluster* similarities.

The *Intracluster* indicator is calculated as the sum of all distances between the several objects in a given cluster (l represents the total number of objects in a cluster) and the mean value (m_i) of the cluster (C_i) in which the object (o_j) reside. The total number of clusters identified in each iteration is represented by t. The *Intracluster* indicator is calculated as follows:

$$Intracluster = \sum_{i=1}^{t} \sum_{j=1}^{l} o_j \in C_i \left| o_j - m_i \right| \tag{3}$$

The *Intercluster* indicator is calculated as the sum of all distances existing between the centres of all the clusters identified in a given iteration. The *Intercluster* indicator is calculated as follows:

$$Intercluster = \sum_{i=1}^{t} \left(\sum_{\substack{j=1 \\ j \neq i}}^{t} \left| m_i - m_j \right| \right) \tag{4}$$

The *ModelQuality* metric can be used by the user in the selection of the appropriate model for a given purpose. However, it is on the minimum value of this metric that we found the STICH outliers model, in the case they exist in the dataset.

For the example provided in section 2.1, $D = \{3, 5, 8, 12, 14, 18, 20, 22, 24\}$, Fig. 3 presents the clusters that emerge from the data, identified by the STICH algorithm.

Five iterations were needed in order to group the nine objects, initially in nine clusters, into one single cluster. Following the *ModelQuality* criterion defined for STICH, the model that equilibrates the difference between the *Intercluster* and *Intracluster* similarity indicators is the model obtained in the 4th iteration, in which the initial data set is divided into two clusters (in this dataset, the presence of outliers is not observed).

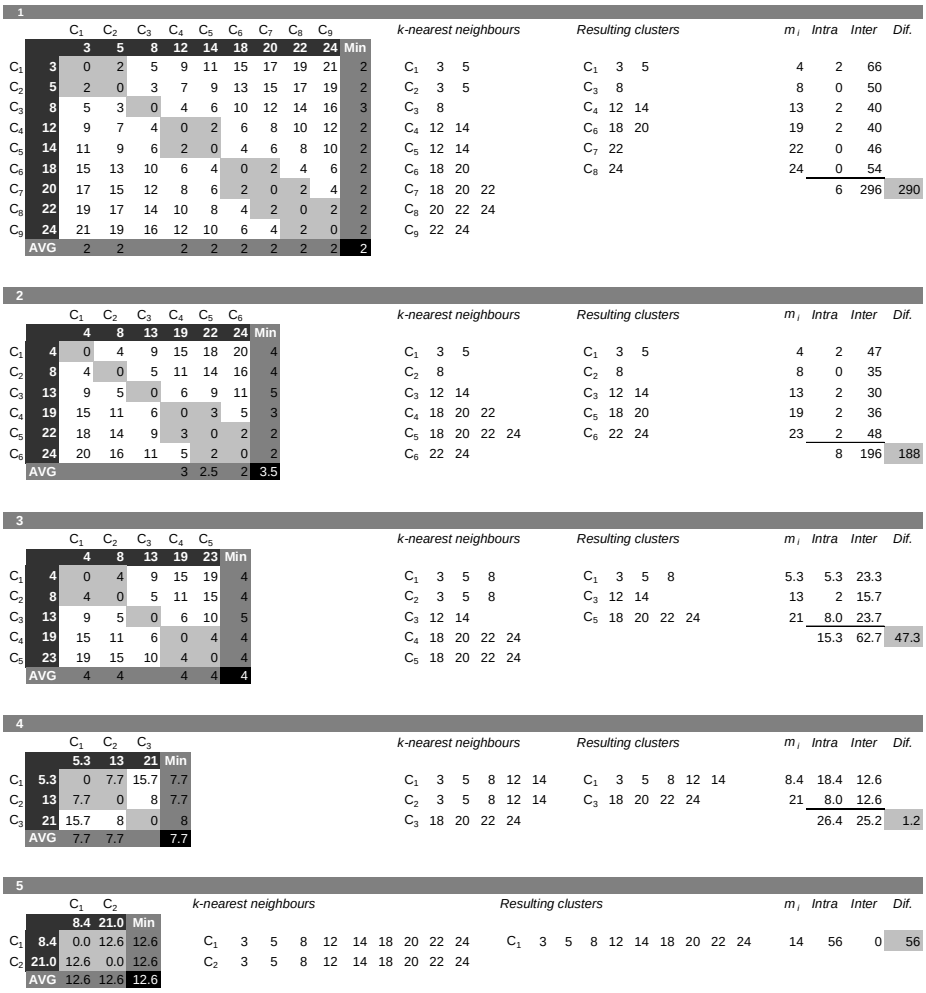

Fig. 3. The STICH approach: an example

3 Results

This section compares the results achieved with the *k-means* and STICH algorithms, in order to identify the differences between the clusters obtained by the two approaches. The two algorithms were used to analyse the same data set: the Population Density attribute for 15 countries of the European Union (a small dataset that allows the summarization of the results achieved in each step of the algorithms).

Table 1 presents the Population Density attribute for the 15 countries. As already mentioned, this data is afterwards analysed by the *k-means* and STICH algorithms.

Table 1. Population Density

Country	Finland	Sweden	Ireland	Spain	Greece
Pop. Density	17.0	21.6	53.9	79.1	82.9

Country	Austria	France	Portugal	Denmark	Luxemburg
Pop. Density	96.7	108.3	111.1	123.9	169.2

Country	Italy	Germany	U. Kingdom	Belgium	Netherlands
Pop. Density	191.7	230.2	240.5	335.9	470.2

3.1 *K-means*

For the analysis of the population density values with the *k-means* algorithm it is necessary to define the number of clusters in which the data will be clustered. As an arbitrary value, lets define $k=3$. As already mentioned, the choice of a value for the parameter k does not emerge from the analysed data – by choosing this value arbitrarily we are imposing some constrains on the achieved results which can be seen as an initial disadvantage of the *k-means* approach. Following the procedure previously described for this algorithm, Table 2 shows the clusters obtained in each of the 8 steps needed by *k-means* for the identification of the final combination, and the respective centroids of the clusters calculated in the iterations. The final result is obtained after 8 iterations, since no changes in the clusters composition are observed from iteration 7 to iteration 8 (stop criterion).

Table 2. Execution of the *k-means* algorithm for the Population Density attribute

Iteration / Indicator value	1	2	3	4	5	6	7	8
17.0	C_1	C_1	C_1	C_1	C_1	C_1	C_1	C_1
21.6	C_2	C_2	C_1	C_1	C_1	C_1	C_1	C_1
53.9	C_3	C_2	C_2	C_1	C_1	C_1	C_1	C_1
79.1	C_3	C_2	C_2	C_2	C_2	C_2	C_2	C_2
82.9	C_3	C_2	C_2	C_2	C_2	C_2	C_2	C_2
96.7	C_3	C_2	C_2	C_2	C_2	C_2	C_2	C_2
108.3	C_3	C_3	C_2	C_3	C_2	C_2	C_2	C_2
111.1	C_3	C_3	C_2	C_3	C_2	C_2	C_2	C_2
123.9	C_3	C_3	C_2	C_3	C_2	C_2	C_2	C_2
169.2	C_3	C_3	C_3	C_3	C_3	C_2	C_2	C_2
191.7	C_3	C_3	C_3	C_3	C_3	C_3	C_2	C_2
230.2	C_3	C_3	C_3	C_3	C_3	C_3	C_3	C_3
240.5	C_3	C_3	C_3	C_3	C_3	C_3	C_3	C_3
335.9	C_3	C_3	C_3	C_3	C_3	C_3	C_3	C_3
470.2	C_3	C_3	C_3	C_3	C_3	C_3	C_3	C_3
Iteration / Average	1	2	3	4	5	6	7	8
C_1	17.0	17.0	19.3	30.8	30.8	30.8	30.8	30.8
C_2	21.6	66.8	93.7	86.3	100.3	110.2	120.4	120.4
C_3	53.9	220.1	273.0	220.1	273.0	293.7	319.2	319.2

As can be noted from the analysis of Table 2, at the end of the process the several regions are grouped as follows:

C_1 = {17, 21.6, 53.9}
C_2 = {79.1, 82.9, 96.7, 108.3, 111.1, 123.9, 169.2, 191.7}
C_3 = {230.2, 240.5, 335.9, 470.2}

3.2 STICH

For the analysis of the population density values with STICH, seven iterations were needed to integrate the initial regions (each one representing a cluster) in a unique cluster (stop criterion). The obtained Space Models, one per iteration, present different levels of aggregation of the regions. In each iteration, the *k-nearest neighbours* of a given region are identified attending to the assumptions defined in section 2.3. The several clusters obtained by STICH, in the analysis of the Population Density indicator, are summarised in Table 3. This table also shows the centroid of each cluster and the difference between the *Intracluster* and *Intercluster* similarity values (*ModelQuality*).

As can be noted in the Table 3, the model for which the *ModelQuality* metric reaches its minimum value is the obtained at the 6th iteration, grouping the initial data into two clusters. This model points out the outlier present in the analysed data, and that is represented by the *470.2*[3] value. The two clusters aggregate the following values:

C_1 = {17, 21.6, 53.9, 79.1, 82.9, 96.7, 108.3, 111.1, 123.9, 169.2, 191.7, 230.2, 240.5, 335.9}
C_2 = {470.2}

Looking at the model obtained by STICH at the 5th iteration, also grouping the data into three clusters (as *k-means*), the obtained clusters are the following:

C_1 = {17, 21.6, 53.9, 79.1, 82.9, 96.7, 108.3, 111.1, 123.9}
C_2 = {169.2, 191.7, 230.2, 240.5, 335.9}
C_3 = {470.2}

In this distribution, the *470.2* value remains alone in the last cluster, denoting its difference to the other values.

[3] Following the Interquartile Range definition [Laurikkala, 2000 #374] for the identification of outliers, it can be confirmed that the value *470.2* is the unique value that represent a possible outlier in the analysed dataset (in this definition, outliers are values exceeding by 1.5 the Interquartile Range below the 25th percentile or above the 75th percentile).

Table 3. Execution of the STICH algorithm for the Population Density attribute

Indicator value \ Iteration	1	2	3	4	5	6	7
17.0	C_1	C_1	C_1	C_1	C_1	C_1	C_1
21.6	C_1	C_1	C_1	C_1	C_1	C_1	C_1
53.9	C_2	C_2	C_2	C_1	C_1	C_1	C_1
79.1	C_3	C_3	C_2	C_1	C_1	C_1	C_1
82.9	C_3	C_3	C_2	C_1	C_1	C_1	C_1
96.7	C_4	C_4	C_2	C_1	C_1	C_1	C_1
108.3	C_5	C_4	C_2	C_1	C_1	C_1	C_1
111.1	C_5	C_4	C_2	C_1	C_1	C_1	C_1
123.9	C_6	C_4	C_2	C_1	C_1	C_1	C_1
169.2	C_7	C_5	C_3	C_2	C_2	C_1	C_1
191.7	C_8	C_5	C_3	C_2	C_2	C_1	C_1
230.2	C_9	C_6	C_4	C_2	C_2	C_1	C_1
240.5	C_9	C_6	C_4	C_2	C_2	C_1	C_1
335.9	C_{10}	C_7	C_5	C_3	C_2	C_1	C_1
470.2	C_{11}	C_8	C_6	C_4	C_3	C_2	C_1
Average \ Iteration	**1**	**2**	**3**	**4**	**5**	**6**	**7**
C_1	19.3	19.3	19.3	77.2	77.2	133.0	155.5
C_2	53.9	53.9	93.7	207.9	233.5	470.2	
C_3	81.0	81.0	180.5	335.9	470.2		
C_4	96.7	110.0	235.4	470.2			
C_5	109.7	180.5	335.9				
C_6	123.9	235.4	470.2				
C_7	169.2	335.9					
C_8	191.7	470.2					
C_9	235.4						
C_{10}	335.9						
C_{11}	470.2						
Model Quality \ Iteration	**1**	**2**	**3**	**4**	**5**	**6**	**7**
$\lvert Intracluster - Intecluster \rvert$	16358.1	10128.1	5904.1	2226.1	1075.1	330.8 *min.*	1409.7

4 Results Evaluation

The results obtained in the previous section by the *k-means* and the STICH algorithms are now analysed, in order to verify the main differences between them.

As it can be noted in Fig. 4, the approach followed by *k-means* identified clusters that are somehow homogeneous in terms of their size, and not clusters that point out characteristics of the analysed data.

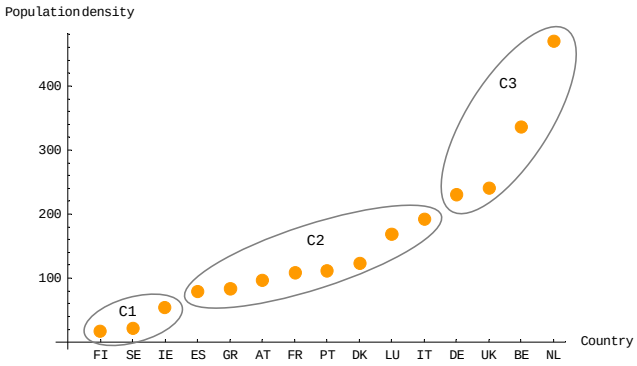

Fig. 4. Division obtained by k-means (*k=3*)

The results obtained by STICH at the 5th iteration, grouping the initial data also into three clusters, and as mentioned earlier, also points out the outlier value present in the dataset. Another characteristic of the results achieved at this point by STICH is the fact that the obtained clusters are not homogeneous in terms of their size, as can be noted in Fig. 5.

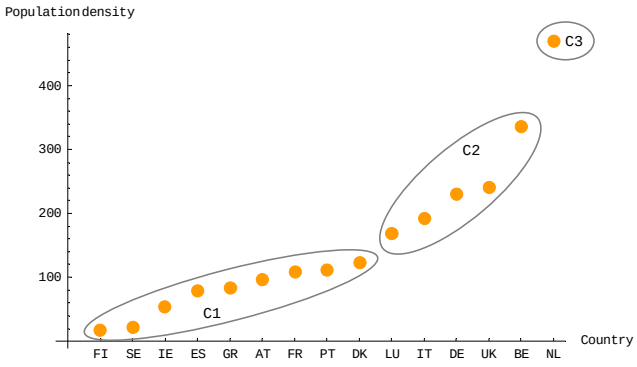

Fig. 5. Division obtained by STICH (*3 clusters*)

Analysing the results obtained by STICH in a graphical way, with the new space geometries created by the algorithm, Fig. 6 and Fig. 7 present the Space Models created at the 2nd (with 8 clusters) and at the 5th (with 3 clusters) iterations, respectively.

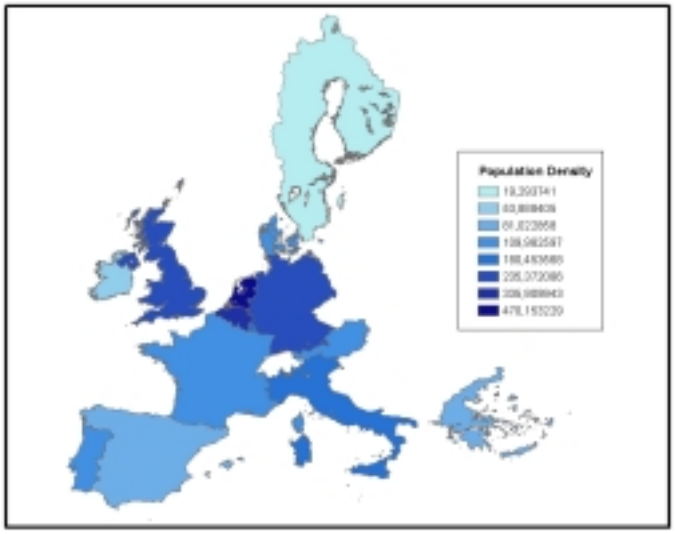

Fig. 6. Space Model obtained by STICH at the 2nd iteration

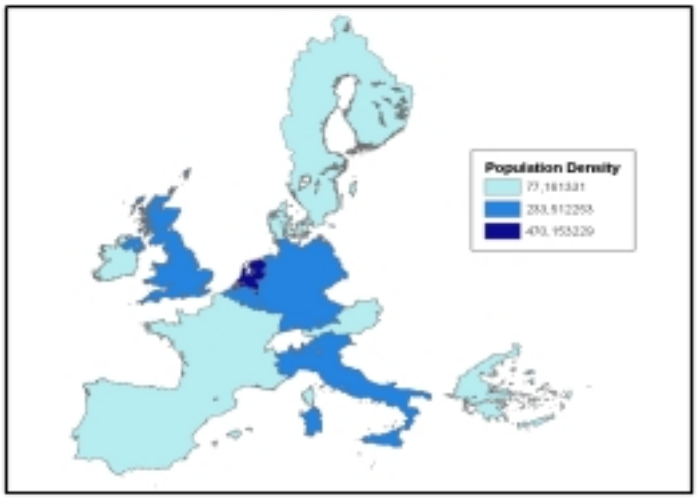

Fig. 7. Space Model obtained by STICH at the 5th iteration

The Space Models created define new space geometries in which elementary regions are grouped accordingly to the similarities between them. These new

divisions of the space can now be used as a tool for data analysis, replacing the traditional maps with administrative subdivisions – sometimes used as consequence of the lack of availability of more appropriate divisions of the space. The Space Models are created imposing no constrains to the divisions that can be achieved, making them more suitable to be used in some specific applications.

As advantages of STICH, and besides the creation of new divisions of the space, we point out:

- The discovery of clusters with arbitrary sizes.
- The avoidance of any previous domain knowledge, as for example the definition of a value for *k*.
- The ability to deal with outliers values, since when they are present in the dataset, the approach undertaken allows their identification.

The advantages of STICH constitute the main disadvantages of the *k-means* algorithm, since it does not permit the identification of clusters with arbitrary shapes, it requires the definition of *k*, and the clusters identified by it are influenced by the outliers present in the dataset (no distinction is made between outliers and non outliers' values).

5 Conclusion

This paper presented STICH, a hierarchical clustering algorithm that allows the creation of Space Models. These models are characterised by the integration of groups of elementary regions that are similar to each other. Clusters of outliers are also formed by STICH, enabling the identification of regions that are very different from the other regions present in the dataset.

The results achieved with STICH were compared with the results achieved by the *k-means* algorithm, a well-known clustering algorithm, allowing the verification of the main differences between these two approaches. While *k-means* finds clusters that are homogeneous in size, STICH identifies clusters of any size. Its main objective is the creation of clusters that naturally emerge from data, imposing no restrictions on the results that can be achieved.

Acknowledgements

This work has been supported by the EPSILON Project (http://get.dsi.uminho.pt/epsilon) funded by the *Information Society Technologies* program from European Commission through contract IST-2001-32389.

References

1. Han, J. and M. Kamber, *Data Mining: Concepts and Techniques*. 2001: Morgan Kaufmann Publishers.
2. Kaufman, L. and P.J. Rousseeuw, *Finding Groups in Data: An Introduction to Cluster Analysis*. 1990: John Wiley & Sons, Inc.
3. Fayyad, U.M., *et al.*, eds. *Advances in Knowledge Discovery and Data Mining*. . 1996, The MIT Press: Massachusetts.
4. Ester, M., *et al.*, *Clustering for Mining in Large Spatial Databases.* KI-Journal, Special Issue on Data Mining, 1998. **1**: p. 18-24.
5. Karypis, G., E.-H. Han, and V. Kumar, *Chameleon: Hierarchical Clustering using Dynamic Modeling.* IEEE Computer, 1999. **32**(8): p. 68-75.
6. MacQueen, J., *Some methods for classification and analysis of multivariate observations*, In Le Cam, L. M. and Neyman, J., editors, Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1, Berkeley, California, University of California Press, 1967, p. 281-297.